CHAPTER 15

15.1. a. Since the regressors are all orthogonal by construction -- $dk_i \cdot dm_i = 0$ for $k \neq m$, and all i -- the coefficient on dm is obtained from the regression y_i on dm_i, $i = 1,...,N$. But this is easily seen to be the fraction of y_i in the sample falling into category m. Therefore, the fitted values are just the cell frequencies, and these are necessarily in [0,1].

b. The fitted values for each category will be the same. If we drop *d1* but add an overall intercept, the overall intercept is the cell frequency for the first category, and the coefficient on *dm* becomes the difference in cell frequency between category m and category one, $m = 2, \ldots, M$.

15.3. a. If $P(y = 1|\mathbf{z}_1, z_2) = \Phi(\mathbf{z}_1\boldsymbol{\delta}_1 + \gamma_1 z_2 + \gamma_2 z_2^2)$ then

$$\frac{\partial P(y = 1|\mathbf{z}_1, z_2)}{\partial z_2} = (\gamma_1 + 2\gamma_2 z_2)\cdot\phi(\mathbf{z}_1\boldsymbol{\delta}_1 + \gamma_1 z_2 + \gamma_2 z_2^2);$$

for given $\mathbf{z}$, this is estimated as

$$(\hat{\gamma}_1 + 2\hat{\gamma}_2 z_2)\cdot\phi(\mathbf{z}_1\hat{\boldsymbol{\delta}}_1 + \hat{\gamma}_1 z_2 + \hat{\gamma}_2 z_2^2),$$

where, of course, the estimates are the probit estimates.

b. In the model

$$P(y = 1|\mathbf{z}_1, z_2, d_1) = \Phi(\mathbf{z}_1\boldsymbol{\delta}_1 + \gamma_1 z_2 + \gamma_2 d_1 + \gamma_3 z_2 d_1),$$

the partial effect of z_2 is

$$\frac{\partial P(y = 1|\mathbf{z}_1, z_2, d_1)}{\partial z_2} = (\gamma_1 + \gamma_3 d_1)\cdot\phi(\mathbf{z}_1\boldsymbol{\delta}_1 + \gamma_1 z_2 + \gamma_2 d_1 + \gamma_3 z_2 d_1).$$

The effect of d_1 is measured as the difference in the probabilities at $d_1 = 1$ and $d_1 = 0$:

$$P(y = 1|\mathbf{z}, d_1 = 1) - P(y = 1|\mathbf{z}, d_1 = 0)$$

$$= \Phi[\mathbf{z}_1\boldsymbol{\delta}_1 + (\gamma_1 + \gamma_3)z_2 + \gamma_2] - \Phi(\mathbf{z}_1\boldsymbol{\delta}_1 + \gamma_1 z_2).$$

Again, to estimate these effects at given $\mathbf{z}$ and -- in the first case, d_1 -- we

just replace the parameters with their probit estimates, and use average or other interesting values of $\mathbf{z}$.

c. We would apply the delta method from Chapter 3. Thus, we would require the full variance matrix of the probit estimates as well as the gradient of the expression of interest, such as $(\gamma_1 + 2\gamma_2 z_2)\cdot\phi(\mathbf{z}_1\boldsymbol{\delta}_1 + \gamma_1 z_2 + \gamma_2 z_2^2)$, with respect to all probit parameters. (Not with respect to the z_j.)

15.5. a. If $P(y = 1|\mathbf{z},q) = \Phi(\mathbf{z}_1\boldsymbol{\delta}_1 + \gamma_1 z_2 q)$ then

$$\frac{\partial P(y = 1|\mathbf{z},q)}{\partial z_2} = \gamma_1 q\cdot\phi(\mathbf{z}_1\boldsymbol{\delta}_1 + \gamma_1 z_2 q),$$

assuming that z_2 is not functionally related to $\mathbf{z}_1$.

b. Write $y^* = \mathbf{z}_1\boldsymbol{\delta}_1 + r$, where $r = \gamma_1 z_2 q + e$, and e is independent of $(\mathbf{z},q)$ with a standard normal distribution. Because q is assumed independent of $\mathbf{z}$, $q|\mathbf{z} \sim \text{Normal}(0,\gamma_1^2 z_2^2 + 1)$; this follows because $E(r|\mathbf{z}) = \gamma_1 z_2 E(q|\mathbf{z}) + E(e|\mathbf{z}) = 0$. Also,

$$\text{Var}(r|\mathbf{z}) = \gamma_1^2 z_2^2 \text{Var}(q|\mathbf{z}) + \text{Var}(e|\mathbf{z}) + 2\gamma_1 z_2 \text{Cov}(q,e|\mathbf{z}) = \gamma_1^2 z_2^2 + 1$$

because $\text{Cov}(q,e|\mathbf{z}) = 0$ by independence between e and $(\mathbf{z},q)$. Thus, $r/\sqrt{\gamma_1^2 z_2^2 + 1}$ has a standard normal distribution independent of $\mathbf{z}$. It follows that

$$P(y = 1|\mathbf{z}) = \Phi\left(\mathbf{z}_1\boldsymbol{\delta}_1/\sqrt{\gamma_1^2 z_2^2 + 1}\right). \tag{15.90}$$

c. Because $P(y = 1|\mathbf{z})$ depends only on γ_1^2, this is what we can estimate along with $\boldsymbol{\delta}_1$. (For example, $\gamma_1 = -2$ and $\gamma_1 = 2$ give exactly the same model for $P(y = 1|\mathbf{z})$.) This is why we define $\rho_1 = \gamma_1^2$. Testing H_0: $\rho_1 = 0$ is most easily done using the score or LM test because, under H_0, we have a standard probit model. Let $\hat{\boldsymbol{\delta}}_1$ denote the probit estimates under the null that $\rho_1 = 0$. Define $\hat{\Phi}_i = \Phi(\mathbf{z}_{i1}\hat{\boldsymbol{\delta}}_1)$, $\hat{\phi}_i = \phi(\mathbf{z}_{i1}\hat{\boldsymbol{\delta}}_1)$, $\hat{u}_i = y_i - \hat{\Phi}_i$, and $\tilde{u}_i \equiv \hat{u}_i/\sqrt{\hat{\Phi}_i(1 - \hat{\Phi}_i)}$

(the standardized residuals). The gradient of the mean function in (15.90) with respect to $\boldsymbol{\delta}_1$, evaluated under the null estimates, is simply $\hat{\phi}_i \mathbf{z}_{i1}$. The only other quantity needed is the gradient with respect to ρ_1 evaluated at the null estimates. But the partial derivative of (15.90) with respect to ρ_1 is, for each i,

$$-(\mathbf{z}_{i1}\boldsymbol{\delta}_1)(z_{i2}^2/2)\left(\rho_1 z_{i2}^2 + 1\right)^{-3/2}\phi\left(\mathbf{z}_{i1}\boldsymbol{\delta}_1/\sqrt{\gamma_1^2 z_{i2}^2 + 1}\right).$$

When we evaluate this at $\rho_1 = 0$ and $\hat{\boldsymbol{\delta}}_1$ we get $-(\mathbf{z}_{i1}\hat{\boldsymbol{\delta}}_1)(z_{i2}^2/2)\hat{\phi}_i$. Then, the score statistic can be obtained as NR_u^2 from the regression

$$\tilde{u}_i \quad \text{on} \quad \hat{\phi}_i \mathbf{z}_{i1}/\sqrt{\hat{\Phi}_i(1 - \hat{\Phi}_i)}, \; (\mathbf{z}_{i1}\hat{\boldsymbol{\delta}}_1) z_{i2}^2 \hat{\phi}_i/\sqrt{\hat{\Phi}_i(1 - \hat{\Phi}_i)};$$

under H_0, $NR_u^2 \overset{a}{\sim} \chi_1^2$.

d. The model can be estimated by MLE using the formulation with ρ_1 in place of γ_1^2. But this is not a standard probit estimation.

15.7. a. The following Stata output is for part a:

```
. reg arr86 pcnv avgsen tottime ptime86 inc86 black hispan born60

      Source |       SS       df       MS              Number of obs =    2725
-------------+------------------------------           F(  8,  2716) =   30.48
       Model |  44.9720916     8  5.62151145            Prob > F      =  0.0000
    Residual |  500.844422  2716  .184405163            R-squared     =  0.0824
-------------+------------------------------           Adj R-squared =  0.0797
       Total |  545.816514  2724   .20037317           Root MSE      =  .42942

------------------------------------------------------------------------------
       arr86 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        pcnv |  -.1543802   .0209336    -7.37   0.000    -.1954275   -.1133329
      avgsen |   .0035024   .0063417     0.55   0.581    -.0089326    .0159374
     tottime |  -.0020613   .0048884    -0.42   0.673    -.0116466     .007524
     ptime86 |  -.0215953   .0044679    -4.83   0.000    -.0303561   -.0128344
       inc86 |  -.0012248    .000127    -9.65   0.000    -.0014738   -.0009759
       black |   .1617183   .0235044     6.88   0.000     .1156299    .2078066
      hispan |   .0892586   .0205592     4.34   0.000     .0489454    .1295718
      born60 |   .0028698   .0171986     0.17   0.867    -.0308539    .0365936
       _cons |   .3609831   .0160927    22.43   0.000      .329428    .3925382
```

```
------------------------------------------------------------------------------

. reg arr86 pcnv avgsen tottime ptime86 inc86 black hispan born60, robust

Regression with robust standard errors                Number of obs =    2725
                                                      F(  8,  2716) =   37.59
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.0824
                                                      Root MSE      =  .42942

------------------------------------------------------------------------------
             |               Robust
       arr86 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        pcnv |  -.1543802    .018964    -8.14   0.000    -.1915656   -.1171948
      avgsen |   .0035024   .0058876     0.59   0.552    -.0080423    .0150471
     tottime |  -.0020613   .0042256    -0.49   0.626     -.010347    .0062244
     ptime86 |  -.0215953   .0027532    -7.84   0.000    -.0269938   -.0161967
       inc86 |  -.0012248   .0001141   -10.73   0.000    -.0014487    -.001001
       black |   .1617183   .0255279     6.33   0.000     .1116622    .2117743
      hispan |   .0892586   .0210689     4.24   0.000     .0479459    .1305714
      born60 |   .0028698   .0171596     0.17   0.867    -.0307774     .036517
       _cons |   .3609831   .0167081    21.61   0.000     .3282214    .3937449
------------------------------------------------------------------------------
```

The estimated effect from increasing *pcnv* from .25 to .75 is about -.154(.5) = -.077, so the probability of arrest falls by about 7.7 points. There are no important differences between the usual and robust standard errors. In fact, in a couple of cases the robust standard errors are notably smaller.

b. The robust statistic and its *p*-value are gotten by using the "test" command after appending "robust" to the regression command:

```
. test avgsen tottime

 ( 1)  avgsen = 0.0
 ( 2)  tottime = 0.0

       F(  2,  2716) =    0.18
            Prob > F =    0.8320


. qui reg arr86 pcnv avgsen tottime ptime86 inc86 black hispan born60

. test avgsen tottime
```

```
 ( 1)  avgsen = 0.0
 ( 2)  tottime = 0.0

       F(  2,  2716) =    0.18
            Prob > F =    0.8360
```

c. The probit model is estimated as follows:

```
. probit arr86 pcnv avgsen tottime ptime86 inc86 black hispan born60

Iteration 0:   log likelihood = -1608.1837
Iteration 1:   log likelihood = -1486.3157
Iteration 2:   log likelihood = -1483.6458
Iteration 3:   log likelihood = -1483.6406

Probit estimates                                  Number of obs   =       2725
                                                  LR chi2(8)      =     249.09
                                                  Prob > chi2     =     0.0000
Log likelihood = -1483.6406                       Pseudo R2       =     0.0774

------------------------------------------------------------------------------
       arr86 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        pcnv |  -.5529248   .0720778    -7.67   0.000    -.6941947   -.4116549
      avgsen |   .0127395   .0212318     0.60   0.548     -.028874    .0543531
     tottime |  -.0076486   .0168844    -0.45   0.651    -.0407414    .0254442
     ptime86 |  -.0812017    .017963    -4.52   0.000    -.1164085   -.0459949
       inc86 |  -.0046346   .0004777    -9.70   0.000    -.0055709   -.0036983
       black |   .4666076   .0719687     6.48   0.000     .3255516    .6076635
      hispan |   .2911005   .0654027     4.45   0.000     .1629135    .4192875
      born60 |   .0112074   .0556843     0.20   0.840    -.0979318    .1203466
       _cons |  -.3138331   .0512999    -6.12   0.000    -.4143791    -.213287
------------------------------------------------------------------------------
```

Now, we must compute the difference in the normal cdf at the two different values of *pcnv*, *black* = 1, *hispan* = 0, *born60* = 1, and at the average values of the remaining variables:

```
. sum avgsen tottime ptime86 inc86

Variable |     Obs        Mean    Std. Dev.       Min        Max
---------+--------------------------------------------------------
  avgsen |    2725    .6322936    3.508031          0       59.2
 tottime |    2725    .8387523    4.607019          0       63.4
 ptime86 |    2725     .387156    1.950051          0         12
   inc86 |    2725    54.96705    66.62721          0        541
```

```
. di -.313 + .0127*.632 - .0076*.839 - .0812*.387 - .0046*54.97 + .467 + .0112
-.1174364

. di normprob(-.553*.75 - .117) - normprob(-.553*.25 - .117)
-.10181543
```

This last command shows that the probability falls by about .10, which is somewhat larger than the effect obtained from the LPM.

d. To obtain the percent correctly predicted for each outcome, we first generate the predicted values of *arr86* as described on page 465:

```
. predict phat
(option p assumed; Pr(arr86))

. gen arr86h = phat > .5

. tab arr86h arr86
```

arr86h	arr86 0	arr86 1	Total
0	1903	677	2580
1	67	78	145
Total	1970	755	2725

```
. di 1903/1970
.96598985

. di 78/755
.10331126
```

For men who were not arrested, the probit predicts correctly about 96.6% of the time. Unfortunately, for the men who were arrested, the probit is correct only about 10.3% of the time. The overall percent correctly predicted is quite high, but we cannot very well predict the outcome we would most like to predict.

e. Adding the quadratic terms gives

```
. probit arr86 pcnv avgsen tottime ptime86 inc86 black hispan born60 pcnvsq
pt86sq inc86sq
```

```
Iteration 0:   log likelihood = -1608.1837
Iteration 1:   log likelihood = -1452.2089
Iteration 2:   log likelihood = -1444.3151
Iteration 3:   log likelihood = -1441.8535
Iteration 4:   log likelihood =  -1440.268
Iteration 5:   log likelihood = -1439.8166
Iteration 6:   log likelihood = -1439.8005
Iteration 7:   log likelihood = -1439.8005

Probit estimates                                  Number of obs   =       2725
                                                  LR chi2(11)     =     336.77
                                                  Prob > chi2     =     0.0000
Log likelihood = -1439.8005                       Pseudo R2       =     0.1047

------------------------------------------------------------------------------
       arr86 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        pcnv |   .2167615   .2604937     0.83   0.405    -.2937968    .7273198
      avgsen |   .0139969   .0244972     0.57   0.568    -.0340166    .0620105
     tottime |  -.0178158   .0199703    -0.89   0.372     -.056957    .0213253
     ptime86 |   .7449712   .1438485     5.18   0.000     .4630333    1.026909
       inc86 |  -.0058786   .0009851    -5.97   0.000    -.0078094   -.0039478
       black |   .4368131   .0733798     5.95   0.000     .2929913     .580635
      hispan |   .2663945    .067082     3.97   0.000     .1349163    .3978727
      born60 |  -.0145223   .0566913    -0.26   0.798    -.1256351    .0965905
      pcnvsq |  -.8570512   .2714575    -3.16   0.002    -1.389098   -.3250042
      pt86sq |  -.1035031   .0224234    -4.62   0.000    -.1474522    -.059554
     inc86sq |   8.75e-06   4.28e-06     2.04   0.041     3.63e-07    .0000171
       _cons |   -.337362   .0562665    -6.00   0.000    -.4476423   -.2270817
------------------------------------------------------------------------------

note: 51 failures and 0 successes completely determined.

. test pcnvsq pt86sq inc86sq

 ( 1)  pcnvsq = 0.0
 ( 2)  pt86sq = 0.0
 ( 3)  inc86sq = 0.0

           chi2(  3) =   38.54
         Prob > chi2 =    0.0000
```

The quadratics are individually and jointly significant. The quadratic in *pcnv* means that, at low levels of *pcnv*, there is actually a positive relationship between probability of arrest and *pcnv*, which does not make much sense. The turning point is easily found as .217/(2*.857) ≈ .127, which means that there is an estimated deterrent effect over most of the range of *pcnv*.

15.9. a. Let $P(y = 1|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$, where $x_1 = 1$. Then, for each i,

$$\ell_i(\boldsymbol{\beta}) = y_i \log(\mathbf{x}_i\boldsymbol{\beta}) + (1 - y_i)\log(1 - \mathbf{x}_i\boldsymbol{\beta}),$$

which is only well-defined for $0 < \mathbf{x}_i\boldsymbol{\beta} < 1$.

b. For any possible estimate $\hat{\boldsymbol{\beta}}$, the log-likelihood function is well-defined only if $0 < \mathbf{x}_i\hat{\boldsymbol{\beta}} < 1$ for all $i = 1,\ldots,N$. Therefore, during the iterations to obtain the MLE, this condition must be checked. It may be impossible to find an estimate that satisfies these inequalities for every observation, especially if N is large.

c. This follows from the KLIC: the true density of y given $\mathbf{x}$ -- evaluated at the true values, of course -- maximizes the KLIC. Since the MLEs are consistent for the unknown parameters, asymptotically the true density will produce the highest average log likelihood function. So, just as we can use an R-squared to choose among different functional forms for $E(y|\mathbf{x})$, we can use values of the log-likelihood to choose among different models for $P(y = 1|\mathbf{x})$ when y is binary.

15.11. We really need to make two assumptions. The first is a conditional independence assumption: given $\mathbf{x}_i = (\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT})$, $(y_{i1},\ldots,y_{iT})$ are independent. This allows us to write

$$f(y_1,\ldots,y_T|\mathbf{x}_i) = f_1(y_1|\mathbf{x}_i)\cdots f_T(y_T|\mathbf{x}_i),$$

that is, the joint density (conditional on $\mathbf{x}_i$) is the product of the marginal densities (each conditional on $\mathbf{x}_i$). The second assumption is a strict exogeneity assumptiond: $D(y_{it}|\mathbf{x}_i) = D(y_{it}|\mathbf{x}_{it})$, $t = 1,\ldots,T$. When we add the standard assumption for pooled probit -- that $D(y_{it}|\mathbf{x}_{it})$ follows a probit model -- then

$$f(y_1,\ldots,y_T|\mathbf{x}_i) = \prod_{t=1}^{T}[G(\mathbf{x}_{it}\boldsymbol{\beta})]^{y_t}[1 - G(\mathbf{x}_{it}\boldsymbol{\beta})]^{1-y_t},$$

and so pooled probit is conditional MLE.

15.13. a. If there are no covariates, there is no point in using any method other than a straight comparison of means. The estimated probabilities for the treatment and control groups, both before and after the policy change, will be identical across models.

b. Let *d2* be a binary indicator for the second time period, and let *dB* be an indicator for the treatment group. Then a probit model to evaluate the treatment effect is

$$P(y = 1|\mathbf{x}) = \Phi(\delta_0 + \delta_1 d2 + \delta_2 dB + \delta_3 d2\cdot dB + \mathbf{x}\boldsymbol{\gamma}),$$

where $\mathbf{x}$ is a vector of covariates. We would estimate all parameters from a probit of y on 1, *d2*, *dB*, *d2*·*dB*, and $\mathbf{x}$ using all observations. Once we have the estimates, we need to compute the "difference-in-differences" estimate, which requires either plugging in a value for $\mathbf{x}$, say $\bar{\mathbf{x}}$, or averaging the differences across $\mathbf{x}_i$. In the former case, we have

$$\begin{aligned}\hat{\theta} \equiv\ & [\Phi(\hat{\delta}_0 + \hat{\delta}_1 + \hat{\delta}_2 + \hat{\delta}_3 + \bar{\mathbf{x}}\hat{\boldsymbol{\gamma}}) - \Phi(\hat{\delta}_0 + \hat{\delta}_2 + \bar{\mathbf{x}}\hat{\boldsymbol{\gamma}})] \\ & - [\Phi(\hat{\delta}_0 + \hat{\delta}_1 + \bar{\mathbf{x}}\hat{\boldsymbol{\gamma}}) - \Phi(\hat{\delta}_0 + \bar{\mathbf{x}}\hat{\boldsymbol{\gamma}})],\end{aligned}$$

and in the latter we have

$$\begin{aligned}\tilde{\theta} \equiv\ & N^{-1}\sum_{i=1}^{N}\{[\Phi(\hat{\delta}_0 + \hat{\delta}_1 + \hat{\delta}_2 + \hat{\delta}_3 + \mathbf{x}_i\hat{\boldsymbol{\gamma}}) - \Phi(\hat{\delta}_0 + \hat{\delta}_2 + \mathbf{x}_i\hat{\boldsymbol{\gamma}})] \\ & - [\Phi(\hat{\delta}_0 + \hat{\delta}_1 + \mathbf{x}_i\hat{\boldsymbol{\gamma}}) - \Phi(\hat{\delta}_0 + \mathbf{x}_i\hat{\boldsymbol{\gamma}})]\}.\end{aligned}$$

Both are estimates of the difference, between groups *B* and *A*, of the change in the response probability over time.

c. We would have to use the delta method to obtain a valid standard error for either $\hat{\theta}$ or $\tilde{\theta}$.

15.15. We should use an interval regression model; equivalently, ordered probit with known cut points. We would be assuming that the underlying GPA is normally distributed conditional on $\mathbf{x}$, but we only observe interval coded data. (Clearly a conditional normal distribution for the GPAs is at best an approximation.) Along with the β_j -- including an intercept -- we estimate σ^2. The estimated coefficients are interpreted as if we had done a linear regression with actual GPAs.

15.17. a. We obtain the joint density by the product rule, since we have independence conditional on $(\mathbf{x},c)$:

$$f(y_1,\ldots,y_G|\mathbf{x},c;\boldsymbol{\gamma}_o) = f_1(y_1|\mathbf{x},c;\boldsymbol{\gamma}_o^1)f_2(y_1|\mathbf{x},c;\boldsymbol{\gamma}_o^2)\cdots f_G(y_G|\mathbf{x},c;\boldsymbol{\gamma}_o^G).$$

b. The density of $(y_1,\ldots,y_G)$ given $\mathbf{x}$ is obtained by integrating out with respect to the distribution of c given $\mathbf{x}$:

$$g(y_1,\ldots,y_G|\mathbf{x};\boldsymbol{\gamma}_o) = \int_{-\infty}^{\infty}\left(\prod_{g=1}^{G} f_g(y_g|\mathbf{x},c;\boldsymbol{\gamma}_o^g)\right)h(c|\mathbf{x};\boldsymbol{\delta}_o)dc,$$

where c is a dummy argument of integration. Because c appears in each $D(y_g|\mathbf{x},c)$, $y_1,\ldots,y_G$ are dependent without conditioning on c.

c. The log likelihood for each i is

$$\log\left[\int_{-\infty}^{\infty}\left(\prod_{g=1}^{G} f_g(y_{ig}|\mathbf{x}_i,c;\boldsymbol{\gamma}^g)\right)h(c|\mathbf{x}_i;\boldsymbol{\delta})dc\right].$$

As expected, this depends only on the observed data, $(\mathbf{x}_i,y_{i1},\ldots,y_{iG})$, and the unknown parameters.

15.19. To be added.

CHAPTER 16

16.1. a. $P[\log(t_i) = \log(c)|\mathbf{x}_i] = P[\log(t_i^*) > \log(c)|\mathbf{x}_i]$

$= P[u_i > \log(c) - \mathbf{x}_i\boldsymbol{\beta}|\mathbf{x}_i] = 1 - \Phi\{[\log(c) - \mathbf{x}_i\boldsymbol{\beta}]/\sigma\}$.

As $c \to \infty$, $\Phi\{[\log(c) - \mathbf{x}_i\boldsymbol{\beta}]/\sigma\} \to 1$, and so $P[\log(t_i) = \log(c)|\mathbf{x}_i] \to 0$ as $c \to \infty$. This simply says that, the longer we wait to censor, the less likely it is that we observe a censored observation.

b. The density of $y_i \equiv \log(t_i)$ (given $\mathbf{x}_i$) when $t_i < c$ is the same as the density of $y_i^* \equiv \log(t_i^*)$, which is just Normal$(\mathbf{x}_i\boldsymbol{\beta},\sigma^2)$. This is because, for $y < \log(c)$, $P(y_i \leq y|\mathbf{x}_i) = P(y_i^* \leq y|\mathbf{x}_i)$. Thus, the density for $y_i = \log(t_i)$ is

$$f(y|\mathbf{x}_i) = 1 - \Phi\{[\log(c) - \mathbf{x}_i\boldsymbol{\beta}]/\sigma\}, \quad y = \log(c)$$

$$f(y|\mathbf{x}_i) = \frac{1}{\sigma}\phi[(y - \mathbf{x}_i\boldsymbol{\beta})/\sigma], \quad y < \log(c).$$

c. $\ell_i(\boldsymbol{\beta},\sigma^2) = 1[y_i = \log(c)]\cdot\log(1 - \Phi\{[\log(c) - \mathbf{x}_i\boldsymbol{\beta}]/\sigma\})$

$+ 1[y_i < \log(c)]\cdot\log\{\sigma^{-1}\phi[(y_i - \mathbf{x}_i\boldsymbol{\beta})/\sigma]\}$.

d. To test H_0: $\boldsymbol{\beta}_2 = \mathbf{0}$, I would probably use the likelihood ratio statistic. This requires estimating the model with all variables, and then the model without $\mathbf{x}_2$. The LR statistic is $\mathcal{LR} = 2(\mathcal{L}_{ur} - \mathcal{L}_r)$. Under H_0, $\mathcal{LR}$ is distributed asymptotically as $\chi^2_{K_2}$.

e. Since u_i is independent of $(\mathbf{x}_i,c_i)$, the density of y_i given $(\mathbf{x}_i,c_i)$ has the same form as the density of y_i given $\mathbf{x}_i$ above, except that c_i replaces c. The assumption that u_i is independent of c_i means that the decision to censor an individual (or other economic unit) is not related to unobservables affecting t_i^*. Thus, in something like an unemployment duration equation, where u_i might contain unobserved ability, we do not wait longer to censor people of lower ability. Note that c_i can be related to $\mathbf{x}_i$. Thus, if $\mathbf{x}_i$ contains something like education, which is treated as exogenous, then the

censoring time can depend on education.

16.3. a. $P(y_i = a_1|\mathbf{x}_i) = P(y_i^* \leq a_1|\mathbf{x}_i) = P[(u_i/\sigma) \leq (a_1 - \mathbf{x}_i\boldsymbol{\beta})/\sigma]$

$= \Phi[(a_1 - \mathbf{x}_i\boldsymbol{\beta})/\sigma]$.

Similarly,

$$P(y_i = a_2|\mathbf{x}_i) = P(y_i^* \geq a_2|\mathbf{x}_i) = P(\mathbf{x}_i\boldsymbol{\beta} + u_i \geq a_2|\mathbf{x}_i)$$
$$= P[(u_i/\sigma) \geq (a_2 - \mathbf{x}_i\boldsymbol{\beta})/\sigma] = 1 - \Phi[(a_2 - \mathbf{x}_i\boldsymbol{\beta})/\sigma]$$
$$= \Phi[-(a_2 - \mathbf{x}_i\boldsymbol{\beta})/\sigma].$$

Next, for $a_1 < y < a_2$, $P(y_i \leq y|\mathbf{x}_i) = P(y_i^* \leq y|\mathbf{x}_i) = \Phi[(y - \mathbf{x}_i\boldsymbol{\beta})/\sigma]$. Taking the derivative of this cdf with respect to y gives the pdf of y_i conditional on $\mathbf{x}_i$ for values of y strictly between a_1 and a_2: $(1/\sigma)\phi[(y - \mathbf{x}_i\boldsymbol{\beta})/\sigma]$.

b. Since $y = y^*$ when $a_1 < y^* < a_2$, $E(y|\mathbf{x}, a_1 < y < a_2) = E(y^*|\mathbf{x}, a_1 < y^* < a_2)$. But $y^* = \mathbf{x}\boldsymbol{\beta} + u$, and $a_1 < y^* < a_2$ if and only if $a_1 - \mathbf{x}\boldsymbol{\beta} < u < a_2 - \mathbf{x}\boldsymbol{\beta}$. Therefore, using the hint,

$$E(y^*|\mathbf{x}, a_1 < y^* < a_2) = \mathbf{x}\boldsymbol{\beta} + E(u|\mathbf{x}, a_1 - \mathbf{x}\boldsymbol{\beta} < u < a_2 - \mathbf{x}\boldsymbol{\beta})$$
$$= \mathbf{x}\boldsymbol{\beta} + \sigma E[(u/\sigma)|\mathbf{x}, (a_1 - \mathbf{x}\boldsymbol{\beta})/\sigma < u/\sigma < (a_2 - \mathbf{x}\boldsymbol{\beta})/\sigma]$$
$$= \mathbf{x}\boldsymbol{\beta} + \sigma\{\phi[(a_1 - \mathbf{x}\boldsymbol{\beta})/\sigma]$$
$$- \phi[(a_2 - \mathbf{x}\boldsymbol{\beta})/\sigma]\}/\{\Phi[(a_2 - \mathbf{x}\boldsymbol{\beta})/\sigma] - \Phi[(a_1 - \mathbf{x}\boldsymbol{\beta})/\sigma]\}$$
$$= E(y|\mathbf{x}, a_1 < y < a_2).$$

Now, we can easily get $E(y|\mathbf{x})$ by using the following:

$$E(y|\mathbf{x}) = a_1P(y = a_1|\mathbf{x}) + E(y|\mathbf{x}, a_1 < y < a_2)\cdot P(a_1 < y < a_2|\mathbf{x})$$
$$+ a_2P(y_2 = a_2|\mathbf{x})$$
$$= a_1\Phi[(a_1 - \mathbf{x}\boldsymbol{\beta})/\sigma]$$
$$+ E(y|\mathbf{x}, a_1 < y < a_2)\cdot\{\Phi[(a_2 - \mathbf{x}\boldsymbol{\beta})/\sigma] - \Phi[(a_1 - \mathbf{x}\boldsymbol{\beta})/\sigma]\}$$
$$+ a_2\Phi[(\mathbf{x}\boldsymbol{\beta} - a_2)/\sigma]$$
$$= a_1\Phi[(a_1 - \mathbf{x}\boldsymbol{\beta})/\sigma]$$

$$+ (\mathbf{x}\boldsymbol{\beta})\cdot\{\Phi[(a_2 - \mathbf{x}\boldsymbol{\beta})/\sigma] - \Phi[(a_1 - \mathbf{x}\boldsymbol{\beta})/\sigma]\} \tag{16.57}$$

$$+ \sigma\{\phi[(a_1 - \mathbf{x}\boldsymbol{\beta})/\sigma] - \phi[(a_2 - \mathbf{x}\boldsymbol{\beta})/\sigma]\}$$

$$+ a_2\Phi[(\mathbf{x}\boldsymbol{\beta} - a_2)/\sigma].$$

c. From part b it is clear that $E(y^* | \mathbf{x}, a_1 < y^* < a_2) \neq \mathbf{x}\boldsymbol{\beta}$, and so it would be a fluke if OLS on the restricted sample consistently estimated $\boldsymbol{\beta}$. The linear regression of y_i on $\mathbf{x}_i$ using only those y_i such that $a_1 < y_i < a_2$ consistently estimates the linear projection of y^* on $\mathbf{x}$ in the subpopulation for which $a_1 < y^* < a_2$. Generally, there is no reason to think that this will have any simple relationship to the parameter vector $\boldsymbol{\beta}$. [In some restrictive cases, the regression on the restricted subsample could consistently estimate $\boldsymbol{\beta}$ up to a common scale coefficient.]

d. We get the log-likelihood immediately from part a:

$$\ell_i(\theta) = 1[y_i = a_1]\log\{\Phi[(a_1 - \mathbf{x}_i\boldsymbol{\beta})/\sigma]\}$$

$$+ 1[y_i = a_2]\log\{\Phi[(\mathbf{x}_i\boldsymbol{\beta} - a_2)/\sigma]\}$$

$$+ 1[a_1 < y_i < a_2]\log\{(1/\sigma)\phi[(y_i - \mathbf{x}_i\boldsymbol{\beta})/\sigma]\}.$$

Note how the indicator function selects out the appropriate density for each of the three possible cases: at the left endpoint, at the right endpoint, or strictly between the endpoints.

e. After obtaining the maximum likelihood estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$, just plug these into the formulas in part b. The expressions can be evaluated at interesting values of $\mathbf{x}$.

f. We can show this by brute-force differentiation of equation (16.57). As a shorthand, write $\phi_1 \equiv \phi[(a_1 - \mathbf{x}\boldsymbol{\beta})/\sigma]$, $\phi_2 \equiv \phi[(a_2 - \mathbf{x}\boldsymbol{\beta})/\sigma] = \phi[(\mathbf{x}\boldsymbol{\beta} - a_2)/\sigma]$, $\Phi_1 \equiv \Phi[(a_1 - \mathbf{x}_i\boldsymbol{\beta})/\sigma]$, and $\Phi_2 \equiv \Phi[(a_2 - \mathbf{x}\boldsymbol{\beta})/\sigma]$. Then

$$\frac{\partial E(y|\mathbf{x})}{\partial x_j} = -(a_1/\sigma)\phi_1\beta_j + (a_2/\sigma)\phi_2\beta_j$$

$$+ (\Phi_2 - \Phi_1)\beta_j + [(\mathbf{x}\boldsymbol{\beta}/\sigma)(\phi_1 - \phi_2)]\beta_j$$

$$+ \{[(a_1 - \mathbf{x}\boldsymbol{\beta})/\sigma]\phi_1\}\beta_j - \{[(a_2 - \mathbf{x}\boldsymbol{\beta})/\sigma]\phi_2\}\beta_j,$$

where the first two parts are the derivatives of the first and third terms, respectively, in (16.57), and the last two lines are obtained from differentiating the second term in $E(y|\mathbf{x})$. Careful inspection shows that all terms cancel except $(\Phi_2 - \Phi_1)\beta_j$, which is the expression we wanted to be left with.

The scale factor is simply the probability that a standard normal random variable falls in the interval $[(a_1 - \mathbf{x}\boldsymbol{\beta})/\sigma, (a_2 - \mathbf{x}\boldsymbol{\beta})/\sigma]$, which is necessarily between zero and one.

g. The partial effects on $E(y|\mathbf{x})$ are given in part f. These are estimated as

$$\{\Phi[(a_2 - \mathbf{x}\hat{\boldsymbol{\beta}})/\hat{\sigma}] - \Phi[(a_1 - \mathbf{x}\hat{\boldsymbol{\beta}})/\hat{\sigma}]\}\hat{\beta}_j, \tag{16.58}$$

where the estimates are the MLEs. We could evaluate these partial effects at, say, $\bar{\mathbf{x}}$. Or, we could average $\{\Phi[(a_2 - \mathbf{x}_i\hat{\boldsymbol{\beta}})/\hat{\sigma}] - \Phi[(a_1 - \mathbf{x}_i\hat{\boldsymbol{\beta}})/\hat{\sigma}]\}$ across all i to obtain the average partial effect. In either case, the scaled $\hat{\beta}_j$ can be compared to the $\hat{\gamma}_j$. Generally, we expect

$$\hat{\gamma}_j \approx \hat{\rho} \cdot \hat{\beta}_j,$$

where $0 < \hat{\rho} < 1$ is the scale factor. Of course, this approximation need not be very good in a partiular application, but it is often roughly true. It does not make sense to directly compare the magnitude of $\hat{\beta}_j$ with that of $\hat{\gamma}_j$. By the way, note that $\hat{\sigma}$ appears in the partial effects along with the $\hat{\beta}_j$; there is no sense in which $\hat{\sigma}$ is "ancillary."

h. For data censoring where the censoring points might change with i, the analysis is essentially the same but a_1 and a_2 are replaced with a_{i1} and a_{i2}. Intepretating the results is even easier, since we act as if we were able to do OLS on an uncensored sample.

16.5. a. The results from OLS estimation of the linear model are

```
. reg hrbens exper age educ tenure married male white nrtheast nrthcen south
union

  Source |       SS       df       MS                  Number of obs =     616
---------+------------------------------                F( 11,   604) =   32.50
   Model |  101.132288    11  9.19384436                Prob > F      =  0.0000
Residual |  170.839786   604  .282847328                R-squared     =  0.3718
---------+------------------------------                Adj R-squared =  0.3604
   Total |  271.972074   615  .442231015                Root MSE      =  .53183

------------------------------------------------------------------------------
  hrbens |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   exper |   .0029862   .0043435      0.688   0.492       -.005544    .0115164
     age |  -.0022495   .0041162     -0.547   0.585      -.0103333    .0058343
    educ |    .082204   .0083783      9.812   0.000       .0657498    .0986582
  tenure |   .0281931   .0035481      7.946   0.000        .021225    .0351612
 married |   .0899016   .0510187      1.762   0.079       -.010294    .1900971
    male |    .251898   .0523598      4.811   0.000       .1490686    .3547274
   white |    .098923   .0746602      1.325   0.186      -.0477021    .2455481
nrtheast |  -.0834306   .0737578     -1.131   0.258      -.2282836    .0614223
 nrthcen |  -.0492621   .0678666     -0.726   0.468      -.1825451     .084021
   south |  -.0284978   .0673714     -0.423   0.672      -.1608084    .1038129
   union |   .3768401   .0499022      7.552   0.000       .2788372    .4748429
   _cons |  -.6999244   .1772515     -3.949   0.000      -1.048028   -.3518203
------------------------------------------------------------------------------
```

b. The Tobit estimates are

```
. tobit hrbens exper age educ tenure married male white nrtheast nrthcen south
union, ll(0)

Tobit Estimates                                   Number of obs =    616
                                                  chi2(11)      = 283.86
                                                  Prob > chi2   = 0.0000
Log Likelihood = -519.66616                       Pseudo R2     = 0.2145

------------------------------------------------------------------------------
  hrbens |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   exper |   .0040631   .0046627      0.871   0.384      -.0050939    .0132201
     age |  -.0025859   .0044362     -0.583   0.560      -.0112981    .0061263
    educ |   .0869168   .0088168      9.858   0.000       .0696015    .1042321
  tenure |   .0287099   .0037237      7.710   0.000        .021397    .0360227
 married |   .1027574   .0538339      1.909   0.057      -.0029666    .2084814
    male |   .2556765   .0551672      4.635   0.000       .1473341     .364019
   white |   .0994408    .078604      1.265   0.206       -.054929    .2538105
```

```
nrtheast |  -.0778461   .0775035     -1.004   0.316      -.2300547    .0743625
 nrthcen |  -.0489422   .0713965     -0.685   0.493      -.1891572    .0912729
   south |  -.0246854   .0709243     -0.348   0.728      -.1639731    .1146022
   union |   .4033519   .0522697      7.717   0.000       .3006999    .5060039
   _cons |  -.8137158   .1880725     -4.327   0.000       -1.18307   -.4443616
---------+------------------------------------------------------------------------
     _se |   .5551027   .0165773              (Ancillary parameter)
------------------------------------------------------------------------------

Obs. summary:        41 left-censored observations at hrbens<=0
                    575 uncensored observations
```

The Tobit and OLS estimates are similar because only 41 of 616 observations, or about 6.7% of the sample, have *hrbens* = 0. As expected, the Tobit estimates are all slightly larger in magnitude; this reflects that the scale factor is always less than unity. Again, the parameter "_se" is $\hat{\sigma}$. You should ignore the phrase "Ancillary parameter" (which essentially means "subordinate") associated with "_se" as it is misleading for corner solution applications: as we know, $\hat{\sigma}^2$ appears directly in $\hat{E}(y|\mathbf{x})$ and $\hat{E}(y|\mathbf{x},y > 0)$.

c. Here is what happens when $exper^2$ and $tenure^2$ are included:

```
. tobit hrbens exper age educ tenure married male white nrtheast nrthcen south
union expersq tenuresq, ll(0)

Tobit Estimates                                   Number of obs =    616
                                                  chi2(13)      = 315.95
                                                  Prob > chi2   = 0.0000
Log Likelihood = -503.62108                       Pseudo R2     = 0.2388

------------------------------------------------------------------------------
  hrbens |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   exper |   .0306652   .0085253      3.597   0.000       .0139224     .047408
     age |  -.0040294   .0043428     -0.928   0.354      -.0125583    .0044995
    educ |   .0802587   .0086957      9.230   0.000       .0631812    .0973362
  tenure |   .0581357   .0104947      5.540   0.000        .037525    .0787463
 married |   .0714831   .0528969      1.351   0.177      -.0324014    .1753675
    male |   .2562597   .0539178      4.753   0.000       .1503703    .3621491
   white |   .0906783   .0768576      1.180   0.239      -.0602628    .2416193
nrtheast |  -.0480194   .0760238     -0.632   0.528       -.197323    .1012841
 nrthcen |   -.033717   .0698213     -0.483   0.629      -.1708394    .1034053
   south |   -.017479   .0693418     -0.252   0.801      -.1536597    .1187017
   union |   .3874497    .051105      7.581   0.000       .2870843    .4878151
 expersq |  -.0005524   .0001487     -3.715   0.000      -.0008445   -.0002604
```

```
tenuresq |  -.0013291   .0004098     -3.243   0.001       -.002134   -.0005242
   _cons |  -.9436572   .1853532     -5.091   0.000      -1.307673   -.5796409
---------+------------------------------------------------------------------
     _se |   .5418171   .0161572              (Ancillary parameter)
------------------------------------------------------------------------------

Obs. summary:        41 left-censored observations at hrbens<=0
                    575 uncensored observations
```

Both squared terms are very signficant, so they should be included in the model.

d. There are nine industries, and we use *ind1* as the base industry:

```
. tobit hrbens exper age educ tenure married male white nrtheast nrthcen south
union expersq tenuresq ind2-ind9, ll(0)

Tobit Estimates                                   Number of obs =    616
                                                  chi2(21)      = 388.99
                                                  Prob > chi2   = 0.0000
Log Likelihood = -467.09766                       Pseudo R2     = 0.2940

------------------------------------------------------------------------------
  hrbens |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   exper |   .0267869   .0081297      3.295   0.001       .0108205    .0427534
     age |  -.0034182   .0041306     -0.828   0.408      -.0115306    .0046942
    educ |   .0789402   .0088598      8.910   0.000         .06154    .0963403
  tenure |    .053115   .0099413      5.343   0.000       .0335907    .0726393
 married |   .0547462   .0501776      1.091   0.276      -.0438005    .1532928
    male |   .2411059   .0556864      4.330   0.000       .1317401    .3504717
   white |   .1188029   .0735678      1.615   0.107      -.0256812    .2632871
nrtheast |  -.1016799   .0721422     -1.409   0.159      -.2433643    .0400045
 nrthcen |  -.0724782   .0667174     -1.086   0.278      -.2035085    .0585521
   south |  -.0379854   .0655859     -0.579   0.563      -.1667934    .0908226
   union |   .3143174   .0506381      6.207   0.000       .2148662    .4137686
 expersq |  -.0004405   .0001417     -3.109   0.002      -.0007188   -.0001623
tenuresq |  -.0013026   .0003863     -3.372   0.000      -.0020613    -.000544
    ind2 |  -.3731778   .3742017     -0.997   0.319      -1.108095    .3617389
    ind3 |  -.0963657    .368639     -0.261   0.794      -.8203574    .6276261
    ind4 |  -.2351539   .3716415     -0.633   0.527      -.9650425    .4947348
    ind5 |   .0209362    .373072      0.056   0.955      -.7117618    .7536342
    ind6 |  -.5083107   .3682535     -1.380   0.168      -1.231545     .214924
    ind7 |   .0033643   .3739442      0.009   0.993      -.7310468    .7377754
    ind8 |  -.6107854    .376006     -1.624   0.105      -1.349246     .127675
    ind9 |  -.3257878   .3669437     -0.888   0.375       -1.04645    .3948746
   _cons |  -.5750527   .4137824     -1.390   0.165      -1.387704    .2375989
---------+--------------------------------------------------------------------
     _se |   .5099298   .0151907              (Ancillary parameter)
------------------------------------------------------------------------------
```

```
Obs. summary:        41 left-censored observations at hrbens<=0
                    575 uncensored observations

. test ind2 ind3 ind4 ind5 ind6 ind7 ind8 ind9

 ( 1)  ind2 = 0.0
 ( 2)  ind3 = 0.0
 ( 3)  ind4 = 0.0
 ( 4)  ind5 = 0.0
 ( 5)  ind6 = 0.0
 ( 6)  ind7 = 0.0
 ( 7)  ind8 = 0.0
 ( 8)  ind9 = 0.0

       F(  8,   595) =    9.66
            Prob > F =    0.0000
```

Each industry dummy variable is individually insignificant at even the 10% level, but the joint Wald test says that they are jointly very significant. This is somewhat unusual for dummy variables that are necessarily orthogonal (so that there is not a multicollinearity problem among them). The likelihood ratio statistic is 2(503.621 - 467.098) = 73.046; notice that this is roughly 8 (= number of restrictions) times the F statistic; the p-value for the LR statistic is also essentially zero. Certainly several estimates on the industry dummies are economically significant, with a worker in, say, industry eight earning about 61 cents less per hour in benefits than comparable worker in industry one. [Remember, in this example, with so few observations at zero, it is roughly legitimate to use the parameter estimates as the partial effects.]

16.7. a. This follows because the densities conditional on $y > 0$ are identical for the Tobit model and Cragg's model. A more general case is done in Section 17.3. Briefly, if $f(\cdot|\mathbf{x})$ is the continuous density of y given $\mathbf{x}$, then the density of y given $\mathbf{x}$ and $y > 0$ is $f(\cdot|\mathbf{x})/[1 - F(0|\mathbf{x})]$, where $F(\cdot|\mathbf{x})$ is the cdf

of y given $\mathbf{x}$. When f is the normal pdf with mean $\mathbf{x}\boldsymbol{\beta}$ and variance σ^2, we get that $f(y|\mathbf{x},y > 0) = \{\Phi(\mathbf{x}\boldsymbol{\beta}/\sigma)\}^{-1}\{\phi[(y - \mathbf{x}\boldsymbol{\beta})/\sigma]/\sigma\}$ for the Tobit model, and this is exactly the density specified for Cragg's model given $y > 0$.

b. From (6.8) we have

$$E(y|\mathbf{x}) = \Phi(\mathbf{x}\boldsymbol{\gamma})\cdot E(y|\mathbf{x},y > 0) = \Phi(\mathbf{x}\boldsymbol{\gamma})[\mathbf{x}\boldsymbol{\beta} + \sigma\lambda(\mathbf{x}\boldsymbol{\beta}/\sigma)].$$

c. This follows very generally -- not just for Cragg's model or the Tobit model -- from (16.8):

$$\log[E(y|\mathbf{x})] = \log[P(y > 0|\mathbf{x})] + \log[E(y|\mathbf{x},y > 0)].$$

If we take the partial derivative with respect to $\log(x_1)$ we clearly get the sum of the elasticities.

16.9. a. A two-limit Tobit model, of the kind analyzed in Problem 16.3, is appropriate, with $a_1 = 0$, $a_2 = 10$.

b. The lower limit at zero is logically necessary considering the kind of response: the smallest percentage of one's income that can be invested in a pension plan is zero. On the other hand, the upper limit of 10 is an arbitrary corner imposed by law. One can imagine that some people at the corner $y = 10$ would choose $y > 10$ if they could. So, we can think of an underlying variable, which would be the percentage invested in the absense of any restrictions. Then, there would be no upper bound required (since we would not have to worry about 100 percent of income being invested in a pension plan).

c. From Problem 16.3(b), with $a_1 = 0$, we have

$$E(y|\mathbf{x}) = (\mathbf{x}\boldsymbol{\beta})\cdot\{\Phi[(a_2 - \mathbf{x}\boldsymbol{\beta})/\sigma] - \Phi(-\mathbf{x}\boldsymbol{\beta}/\sigma)\}$$
$$+ \sigma\{\phi(\mathbf{x}\boldsymbol{\beta}/\sigma) - \phi[(a_2 - \mathbf{x}\boldsymbol{\beta})/\sigma]\} + a_2\Phi[(\mathbf{x}\boldsymbol{\beta} - a_2)/\sigma].$$

Taking the derivative of this function with respect to a_2 gives

$$\partial E(y|\mathbf{x})/\partial a_2 = (\mathbf{x}\boldsymbol{\beta}/\sigma)\cdot\phi[(a_2 - \mathbf{x}\boldsymbol{\beta})/\sigma] + [(a_2 - \mathbf{x}\boldsymbol{\beta})/\sigma]\cdot\phi[(a_2 - \mathbf{x}\boldsymbol{\beta})/\sigma]$$
$$+ \Phi[(\mathbf{x}\boldsymbol{\beta} - a_2)/\sigma] - (a_2/\sigma)\phi[(\mathbf{x}\boldsymbol{\beta} - a_2)/\sigma]$$
$$= \Phi[(\mathbf{x}\boldsymbol{\beta} - a_2)/\sigma]. \qquad (16.59)$$

We can plug in $a_2 = 10$ to obtain the approximate effect of increasing the cap from 10 to 11. For a given value of $\mathbf{x}$, we would compute $\Phi[(\mathbf{x}\hat{\boldsymbol{\beta}} - 10)/\hat{\sigma}]$, where $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$ are the MLEs. We might evaluate this expression at the sample average of $\mathbf{x}$ or at other interesting values (such as across gender or race).

d. If $y_i < 10$ for $i = 1,\ldots,N$, $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$ are just the usual Tobit estimates with the "censoring" at zero.

16.11. No. OLS always consistently estimates the parameters of a linear projection -- provided the second moments of y and the x_j are finite, and Var($\mathbf{x}$) has full rank K -- regardless of the nature of y or $\mathbf{x}$. That is why a linear regression analysis is always a reasonable first step for binary outcomes, corner solution outcomes, and count outcomes, provided there is not true data censoring.

16.13. This extension has no practical effect on how we estimate an unobserved effects Tobit or probit model, or how we estimate a variety of unobserved effects panel data models with conditional normal heterogeneity. We simply have

$$c_i = -\left(T^{-1}\sum_{t=1}^{T}\boldsymbol{\pi}_t\right)\boldsymbol{\xi} + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i \equiv \psi + \ddot{\mathbf{x}}_i\boldsymbol{\xi} + a_i,$$

where $\psi \equiv -\left(T^{-1}\sum_{t=1}^{T}\boldsymbol{\pi}_t\right)\boldsymbol{\xi}$. Of course, any aggregate time dummies explicitly get swept out of $\ddot{\mathbf{x}}_i$ in this case but would usually be included in $\mathbf{x}_{it}$.

An interesting follow-up question would have been: What if we standardize each $\mathbf{x}_{it}$ by its cross-sectional mean *and* variance at time t, and

assume c_i is related to the mean and variance of the standardized vectors. In other words, let $\mathbf{z}_{it} \equiv (\mathbf{x}_{it} - \boldsymbol{\pi}_t)\boldsymbol{\Omega}_t^{-1/2}$, $t = 1,\ldots,T$, for each random draw i from the population. Then, we might assume $c_i|\mathbf{x}_i \sim \text{Normal}(\psi + \bar{\mathbf{z}}_i\boldsymbol{\xi},\sigma_a^2)$ (where, again, $\mathbf{z}_{it}$ would not contain aggregate time dummies). This is the kind of scenario that is handled by Chamberlain's more general assumption concerning the relationship between c_i and $\mathbf{x}_i$: $c_i = \psi + \sum_{r=1}^{T}\mathbf{x}_{ir}\boldsymbol{\lambda}_r + a_i$, where $\boldsymbol{\lambda}_r = \boldsymbol{\Omega}_r^{-1/2}\boldsymbol{\xi}/T$, $t = 1,2,\ldots,T$. Alternatively, one could estimate estimate $\boldsymbol{\pi}_t$ and $\boldsymbol{\Omega}_t$ for each t using the cross section observations $\{\mathbf{x}_{it}: i = 1,2,\ldots,N\}$. The usual sample means and sample variance matrices, say $\hat{\boldsymbol{\pi}}_t$ and $\hat{\boldsymbol{\Omega}}_t$, are consistent and $\sqrt{N}$–asymptotically normal. Then, form $\hat{\mathbf{z}}_{it} \equiv \hat{\boldsymbol{\Omega}}_t^{-1/2}(\mathbf{x}_{it} - \hat{\boldsymbol{\pi}}_t)$, and proceed with the usual Tobit (or probit) unobserved effects analysis that includes the time averages $\bar{\hat{\mathbf{z}}}_i = T^{-1}\sum_{t=1}^{T}\hat{\mathbf{z}}_{it}$. This is a rather simple two-step estimation method, but accounting for the sample variation in $\hat{\boldsymbol{\pi}}_t$ and $\hat{\boldsymbol{\Omega}}_t$ would be cumbersome. It may be possible to use a much larger to obtain $\hat{\boldsymbol{\pi}}_t$ and $\hat{\boldsymbol{\Omega}}_t$, in which case one might ignore the sampling error in the first-stage estimates.

16.15. To be added.

CHAPTER 17

17.1. If you are interested in the effects of things like age of the building and neighborhood demographics on fire damage, given that a fire has occured, then there is no problem. We simply need a random sample of buildings that actually caught on fire. You might want to supplement this with an analysis of the probability that buildings catch fire, given building and neighborhood characteristics. But then a two-stage analysis is appropriate.

17.3. This is essentially given in equation (17.14). Let y_i given $\mathbf{x}_i$ have density $f(y|\mathbf{x}_i,\boldsymbol{\beta},\boldsymbol{\gamma})$, where $\boldsymbol{\beta}$ is the vector indexing $E(y_i|\mathbf{x}_i)$ and $\boldsymbol{\gamma}$ is another set of parameters (usually a single variance parameter). Then the density of y_i given $\mathbf{x}_i$, $s_i = 1$, when $s_i = 1[a_1(\mathbf{x}_i) < y_i < a_2(\mathbf{x}_i)]$, is

$$p(y|\mathbf{x}_i,s_i=1) = \frac{f(y|\mathbf{x}_i;\boldsymbol{\beta},\boldsymbol{\gamma})}{F(a_2(\mathbf{x}_i)|\mathbf{x}_i;\boldsymbol{\beta},\boldsymbol{\gamma}) - F(a_1(\mathbf{x}_i)|\mathbf{x}_i;\boldsymbol{\beta},\boldsymbol{\gamma})}, \quad a_1(\mathbf{x}_i) < y < a_2(\mathbf{x}_i).$$

In the Hausman and Wise (1977) study, $y_i = \log(income_i)$, $a_1(\mathbf{x}_i) = -\infty$, and $a_2(\mathbf{x}_i)$ was a function of family size (which determines the official poverty level).

17.5. If we replace y_2 with $\hat{y}_2$, we need to see what happens when $y_2 = \mathbf{z}\boldsymbol{\delta}_2 + v_2$ is plugged into the structural mode:

$$\begin{aligned} y_1 &= \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1\cdot(\mathbf{z}\boldsymbol{\delta}_2 + v_2) + u_1 \\ &= \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1\cdot(\mathbf{z}\boldsymbol{\delta}_2) + (u_1 + \alpha_1 v_2). \end{aligned} \tag{17.81}$$

So, the procedure is to replace $\boldsymbol{\delta}_2$ in (17.81) its $\sqrt{N}$-consistent estimator, $\hat{\boldsymbol{\delta}}_2$. The key is to note how the error term in (17.81) is $u_1 + \alpha_1 v_2$. If the selection correction is going to work, we need the expected value of $u_1 + \alpha_1 v_2$ given $(\mathbf{z}, v_3)$ to be linear in v_3 (in particular, it cannot depend on $\mathbf{z}$). Then we can write

$$E(y_1|\mathbf{z}, v_3) = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1\cdot(\mathbf{z}\boldsymbol{\delta}_2) + \gamma_1 v_3,$$

where $E[(u_1 + \alpha_1 v_2)|v_3] = \gamma_1 v_3$ by normality. Conditioning on $y_3 = 1$ gives

$$E(y_1|\mathbf{z}, y_3 = 1) = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1\cdot(\mathbf{z}\boldsymbol{\delta}_2) + \gamma_1\lambda(\mathbf{z}\boldsymbol{\delta}_3). \tag{17.82}$$

A sufficient condition for (17.82) is that (u_1, v_2, v_3) is independent of $\mathbf{z}$ with a trivariate normal distribution. We can get by with less than this, but the nature of v_2 is restricted. If we use an IV approach, we need assume

nothing about v_2 except for the usual linear projection assumption.

As a practical matter, if we cannot write $y_2 = \mathbf{z}\boldsymbol{\delta}_2 + v_2$, where v_2 is independent of $\mathbf{z}$ and approximately normal, then the OLS alternative will not be consistent. Thus, equations where y_2 is binary, or is some other variable that exhibits nonnormality, cannot be consistently estimated using the OLS procedure. This is why 2SLS is generally preferred.

17.7. a. Substitute the reduced forms for y_1 and y_2 into the third equation:

$$y_3 = \max(0,\alpha_1(\mathbf{z}\boldsymbol{\delta}_1) + \alpha_2(\mathbf{z}\boldsymbol{\delta}_2) + \mathbf{z}_3\boldsymbol{\delta}_3 + v_3)$$

$$\equiv \max(0,\mathbf{z}\boldsymbol{\pi}_3 + v_3),$$

where $v_3 \equiv u_3 + \alpha_1 v_1 + \alpha_2 v_2$. Under the assumptions given, v_3 is indepdent of $\mathbf{z}$ and normally distributed. Thus, if we knew $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$, we could consistently estimate α_1, α_2, and $\boldsymbol{\delta}_3$ from a Tobit of y_3 on $\mathbf{z}\boldsymbol{\delta}_1$, $\mathbf{z}\boldsymbol{\delta}_2$, and $\mathbf{z}_3$. >From the usual argument, consistent estimators are obtained by using initial consistent estimators of $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$. Estimation of $\boldsymbol{\delta}_2$ is simple: just use OLS using the entire sample. Estimation of $\boldsymbol{\delta}_1$ follows exactly as in Procedure 17.3 using the system

$$y_1 = \mathbf{z}\boldsymbol{\delta}_1 + v_1 \tag{17.83}$$

$$y_3 = \max(0,\mathbf{z}\boldsymbol{\pi}_3 + v_3), \tag{17.84}$$

where y_1 is observed only when $y_3 > 0$.

Given $\hat{\boldsymbol{\delta}}_1$ and $\hat{\boldsymbol{\delta}}_2$, form $\mathbf{z}_i\hat{\boldsymbol{\delta}}_1$ and $\mathbf{z}_i\hat{\boldsymbol{\delta}}_2$ for each observation i in the sample. Then, obtain $\hat{\alpha}_1$, $\hat{\alpha}_2$, and $\hat{\boldsymbol{\delta}}_3$ from the Tobit

$$y_{i3} \text{ on } (\mathbf{z}_i\hat{\boldsymbol{\delta}}_1), (\mathbf{z}_i\hat{\boldsymbol{\delta}}_2), \mathbf{z}_{i3}$$

using all observations.

For identification, $(\mathbf{z}\boldsymbol{\delta}_1,\mathbf{z}\boldsymbol{\delta}_2,\mathbf{z}_3)$ can contain no exact linear dependencies. Necessary is that there must be at least two elements in $\mathbf{z}$ not

also in $\mathbf{z}_3$.

Obtaining the correct asymptotic variance matrix is complicated. It is most easily done in a generalized method of moments framework.

b. This is not very different from part a. The only difference is that $\boldsymbol{\delta}_2$ must be estimated using Procedure 17.3. Then follow the steps from part a.

c. We need to estimate the variance of u_3, σ_3^2.

17.9. To be added.

17.11. a. There is no sample selection problem because, by definition, you have specified the distribution of y given $\mathbf{x}$ and $y > 0$. We only need to obtain a random sample from the subpopulation with $y > 0$.

b. Again, there is no sample selection bias because we have specified the conditional expectation for the population of interest. If we have a random sample from that population, NLS is generally consistent and $\sqrt{N}$–asymptotically normal.

c. We would use a standard probit model. Let $w = 1[y > 0]$. Then w given $\mathbf{x}$ follows a probit model with $P(w = 1|\mathbf{x}) = \Phi(\mathbf{x}\boldsymbol{\gamma})$.

d. $E(y|\mathbf{x}) = P(y > 0|\mathbf{x}) \cdot E(y|\mathbf{x}, y > 0) = \Phi(\mathbf{x}\boldsymbol{\gamma}) \cdot \exp(\mathbf{x}\boldsymbol{\beta})$. So we would plug in the NLS estimator of $\boldsymbol{\beta}$ and the probit estimator of $\boldsymbol{\gamma}$.

e. Not when you specify the conditional distributions, or conditional means, for the two parts. By definition, there is no sample selection problem. Confusion arises, I think, when two part models are specified with unobservables that may be correlated. For example, we could write

$$y = w \cdot \exp(\mathbf{x}\boldsymbol{\beta} + u),$$

$$w = 1[\mathbf{x}\boldsymbol{\gamma} + v > 0],$$

so that $w = 0 \Rightarrow y = 0$. Assume that (u,v) is independent of $\mathbf{x}$. Then, if u and v are independent -- so that u is independent of $(\mathbf{x},w)$ -- we have

$$E(y|\mathbf{x},w) = w\cdot\exp(\mathbf{x}\boldsymbol{\beta})E[\exp(u)|\mathbf{x},w] = w\cdot\exp(\mathbf{x}\boldsymbol{\beta})E[\exp(u)],$$

which implies the specification in part b (by setting $w = 1$, once we absorb $E[\exp(u)]$ into the intercept). The interesting twist here is if u and v are correlated. Given $w = 1$, we can write $\log(y) = \mathbf{x}\boldsymbol{\beta} + u$. So

$$E[\log(y)|\mathbf{x},w = 1] = \mathbf{x}\boldsymbol{\beta} + E(u|\mathbf{x},w = 1).$$

If we make the usual linearity assumption, $E(u|v) = \rho v$ and assume a standard normal distribution for v then we have the usual inverse Mills ratio added to the linear model:

$$E[\log(y)|\mathbf{x},w = 1] = \mathbf{x}\boldsymbol{\beta} + \rho\lambda(\mathbf{x}\boldsymbol{\gamma}).$$

A two-step strategy for estimating $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ is pretty clear. First, estimate a probit of w_i on $\mathbf{x}_i$ to get $\hat{\boldsymbol{\gamma}}$ and $\lambda(\mathbf{x}_i\hat{\boldsymbol{\gamma}})$. Then, using the $y_i > 0$ observations, run the regression $\log(y_i)$ on $\mathbf{x}_i$, $\lambda(\mathbf{x}_i\hat{\boldsymbol{\gamma}})$ to obtain $\hat{\boldsymbol{\beta}}$, $\hat{\rho}$. A standard t statistic on $\hat{\rho}$ is a simple test of $\text{Cov}(u,v) = 0$.

This two-step procedure reveals a potential problem with the model that allows u and v to be correlated: adding the inverse Mills ratio means that we are adding a nonlinear function of $\mathbf{x}$. In other words, identification of $\boldsymbol{\beta}$ comes entirely from the nonlinearity of the IMR, which we warned about in this chapter. Ideally, we would have a variable that affects $P(w = 1|\mathbf{x})$ that can be excluded from $\mathbf{x}\boldsymbol{\beta}$. In labor economics, where two-part models are used to allow for fixed costs of entering the labor market, one would try to find a variable that affects the fixed costs of being employed that does not affect the choice of hours.

If we assume (u,v) is multivariate normal, with mean zero, then we can use a full maximum likelihood procedure. While this would be a little less

robust, making full distributional assumptions has a subtle advantage: we can then compute partial effects on $E(y|\mathbf{x})$ and $E(y|\mathbf{x},y > 0)$. Even with a full set of assumptions, the partial effects are not straightforward to obtain. For one,

$$E(y|\mathbf{x},y > 0) = \exp(\mathbf{x},\boldsymbol{\beta})\cdot E[\exp(u)|\mathbf{x},w = 1)],$$

where $E[\exp(u)|\mathbf{x},w = 1)]$ can be obtained under joint normality. A similar example is given in Section 19.5.2; see, particularly, equation (19.44). Then, we can multiply this expectation by $P(w = 1|\mathbf{x}) = \Phi(\mathbf{x}\boldsymbol{\gamma})$. The point is that we cannot simply look at $\boldsymbol{\beta}$ to obtain partial effects of interest. This is very different from the sample selection model.

17.13. a. We cannot use censored Tobit because that requires observing $\mathbf{x}$ when whatever the value of y. Instead, we can use truncated Tobit: we use the distribution of y given $\mathbf{x}$ and $y > 0$. Notice that our reason for using truncated Tobit differs from the usual application. Usually, the underlying variable y of interest has a conditional normal distribution in the population. Here, y given $\mathbf{x}$ follows a standard Tobit model in the population (for a corner solution outcome).

b. Provided $\mathbf{x}$ varies enough in the subpopulation where $y > 0$ such that b. rank $E(\mathbf{x}'\mathbf{x}|y > 0) = K$, the parameters. In the case where an element of $\mathbf{x}$ is a derived price, we need sufficient price variation for the population that consumes some of the good. Given such variation, we can estimate $E(y|\mathbf{x}) = \Phi(\mathbf{x}\boldsymbol{\beta}/\sigma)\mathbf{x}\boldsymbol{\beta} + \sigma\phi(\mathbf{x}\boldsymbol{\beta}/\sigma)$ because we have made the assumption that y given $\mathbf{x}$ follows a Tobit in the full population.